SLDS Issue Brief

Centralized vs. Federated: System Models for P-20W+ Data Systems

States' needs for information about the progress of students, schools, districts, and programs continue to expand beyond the boundaries of K12 education to encompass the broader spectrum of P-20W+ information. To meet this need, data must be brought together from multiple sources across agencies and organizations. States have approached the challenge of maintaining and providing secure access to data linked across organizations through two predominate data models, each with its own challenges and advantages. The centralized model for P-20W+ (early childhood through workforce) statewide longitudinal data systems (SLDSs) is a single, integrated data repository that contains, maintains, and provides secure access to data from all participating agencies and organizations. The federated P-20W+ model temporarily links data from repositories maintained by participating organizations to create a report or to generate a dataset.

This document is intended to help state agencies through the process of determining whether a centralized or federated model (or a hybrid approach which includes aspects of both models) will best suit their environment and stakeholder needs. This issue brief will address key questions that should be considered early in the development or restructuring of a P-20W+ system and describes how federated and centralized models bring together data from agencies across a state's P-20W+ environment and make those data useful for and accessible to education stakeholders.

Key Questions to Consider

A clear understanding of your state's unique environment will inform decisions about your system's development and will improve the likelihood that it will meet your end users' information needs. Whether you implement a centralized or a federated system, —or combine some aspects of both approaches—all agencies must address certain fundamental questions and issues. For example, regardless of the data system model, a solid interagency data governance program is required to establish the clear roles, responsibilities, and ownership that are critical to the success of a P-20W+ SLDS.

The following issues, many of which involve aspects of SLDS development beyond system design, should be considered early in P-20W+ system planning:

1. **State policy/legislation**. What are your state's policies regarding data consolidation and exchange? Have there been changes to statutes since your data systems were first developed, or new legislation that limits or facilitates exchange of data between state agencies? For example, are there regulations that limit your state's ability to maintain linked data across agencies? Does any legislation mandate the development of a certain data system model?
2. **Stakeholder information needs**. What P-20W+ longitudinal data do your stakeholders need to inform policy and evaluate programs? Do you need a system

P-20W+ refers to data from prekindergarten (early childhood), K12, and postsecondary through postgraduate education, along with workforce and other outcomes data (e.g., public assistance and corrections data). The specific agencies and other organizations that participate in the P-20W+ initiative vary from state to state.

This product of the Institute of Education Sciences (IES) SLDS Grant Program was developed with the help of knowledgeable staff from state education agencies and partner organizations. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program.

For more information on the IES SLDS Grant Program or for support with system development, please visit http://nces.ed.gov/programs/SLDS.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

to support compliance requirements, respond to data requests from researchers, and produce standard reports on a scheduled basis, or one that can support a broader and less structured array of users and uses?

3. **Governance**. Will a single agency own and host the system, or will ownership be shared among contributing agencies? Do agencies and organizations in your state adhere to a common data standard such as the Common Education Data Standards (CEDS)? Would all participating agencies agree to a common set of rules, or would the agencies require their own rules that would need to be aligned? Can statewide data verification and cleansing processes be implemented to ensure high quality and consistency? Do you have a process for reliably matching records across systems and for reconciling discrepancies that are identified?
4. **Funding**. What funding is available for the development, implementation, and maintenance of the P-20W+ system? Is there a single organization or agency that will serve as the fiscal agent for the system?
5. **Sustainability and responsibility**. How will resources be acquired and allocated for ongoing support, hosting, and maintenance? Will your existing resources be sufficient to support the system over time, or will additional staff and funding be needed? If you are currently using grant funding to develop your SLDS, how will your state sustain the SLDS after that funding is exhausted? What agency, agencies, or organization will be assigned or assume responsibility for maintaining the system over the long term? Are resources available for enhancements needed to ensure the system remains relevant to stakeholder needs?
6. **Staffing capacity**. What are the staffing resources available from participating agencies? Will the work be allocated across participating agencies and organizations, or are there staff members in a single agency/organization that will be dedicated to managing and maintaining the system?
7. **Timeline**. What is your timeline for implementation? Do you have a single deadline to complete all the integration work, or will you look for quick wins and incorporate data from separate source systems over time? How will work be prioritized? Will you be responding to specific immediate data needs or will there be a planned phased approach?
8. **Scalability**. How scalable must your system be? Will there be a need for the system to accommodate other data sources after the initial implementation?
9. **Data sharing culture**. How do you and your partner agencies approach data ownership and sharing? Are all the participating agencies and organizations open to sharing datasets (using appropriate privacy processes) with researchers and other outside entities, or do the participating partners have different stances on this?
10. **Privacy protection**. Are there federal and state laws that affect interagency data sharing in your state? What are the participating agencies' responsibilities around governance and the protection of combined datasets? Are the datasets being shared truly de-identified?[1] Will the data be subject to requirements of the Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA) or other laws? Many state laws require memoranda of understanding or contracts for multi-agency data sharing. How will these requirements be met, and how will they affect the work over time? Is there a limit to the number of years that data can be shared or used?

Centralized and Federated P-20W+ Models: What Are They, and How Do They Compare?

Centralized and federated P-20W+ SLDSs share basic characteristics in terms of data sources and the ultimate presentation of data and information to users. However, there are several key differences related to whether and how data are combined, stored, and accessed.

In a **centralized data system**, data from participating source systems are copied to a single, centrally located data repository where they are organized, integrated, and stored using a common data standard. Data sharing agreements or memoranda of understanding (MOUs) specify the data to be shared and how those data will be managed and used. As depicted in figure 1 (page 3), data in a centralized P-20W+ SLDS are periodically matched, integrated, and loaded into a central repository. Users query the system and can access the data that they have been authorized to view and use.

In a **federated data system**, individual source systems maintain control over their own data but agree through MOUs or data sharing agreements to share the data with other participating systems upon request. As depicted in figure 2 (page 3), data are queried from the independent source systems and records are linked to fulfill a data requestor's information needs. The linked data are not stored by the system, but rather are cached, delivered to the requestor, and then removed.

1 De-identification of data refers to the process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. Although it might not be possible to remove the disclosure risk completely, de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual. De-identified data may be shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, general public, and researchers (34 CFR §99.31(b)(1)).

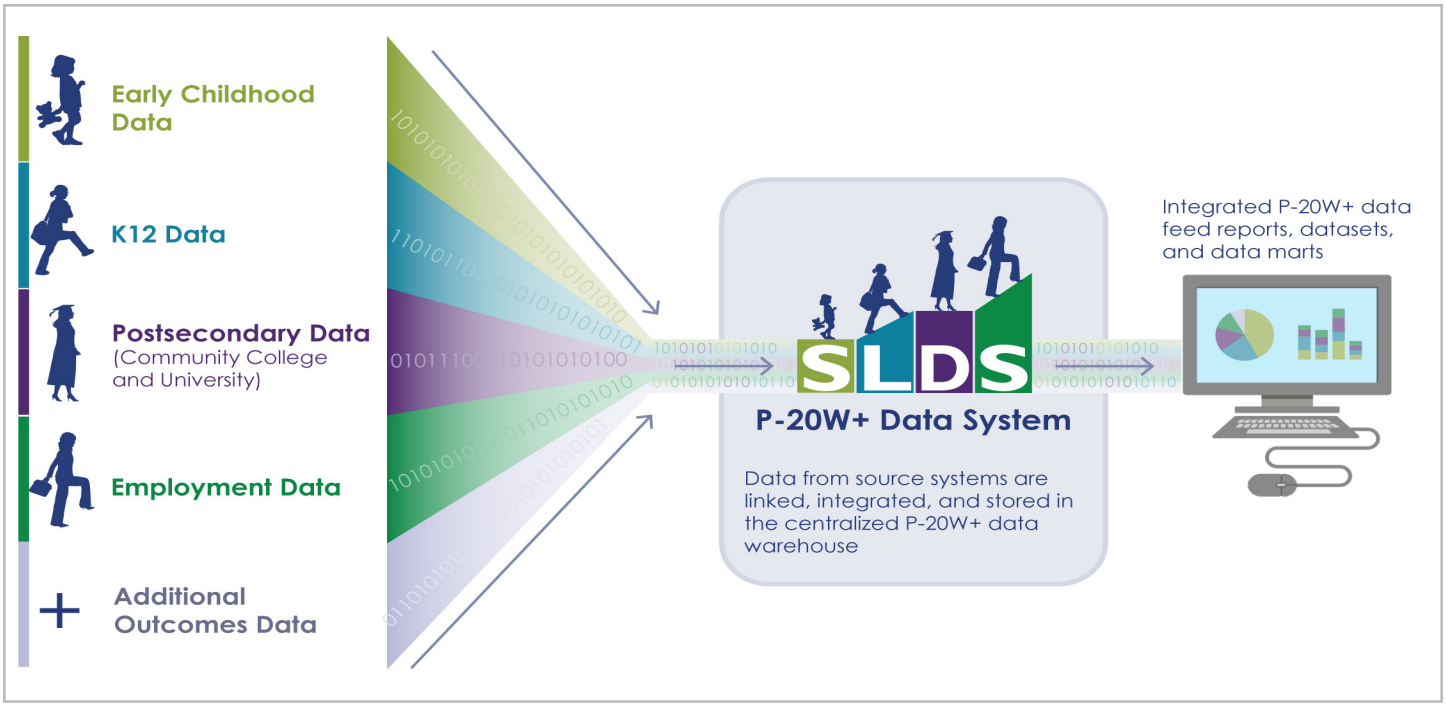

Figure 1. Basic structure of a centralized data system

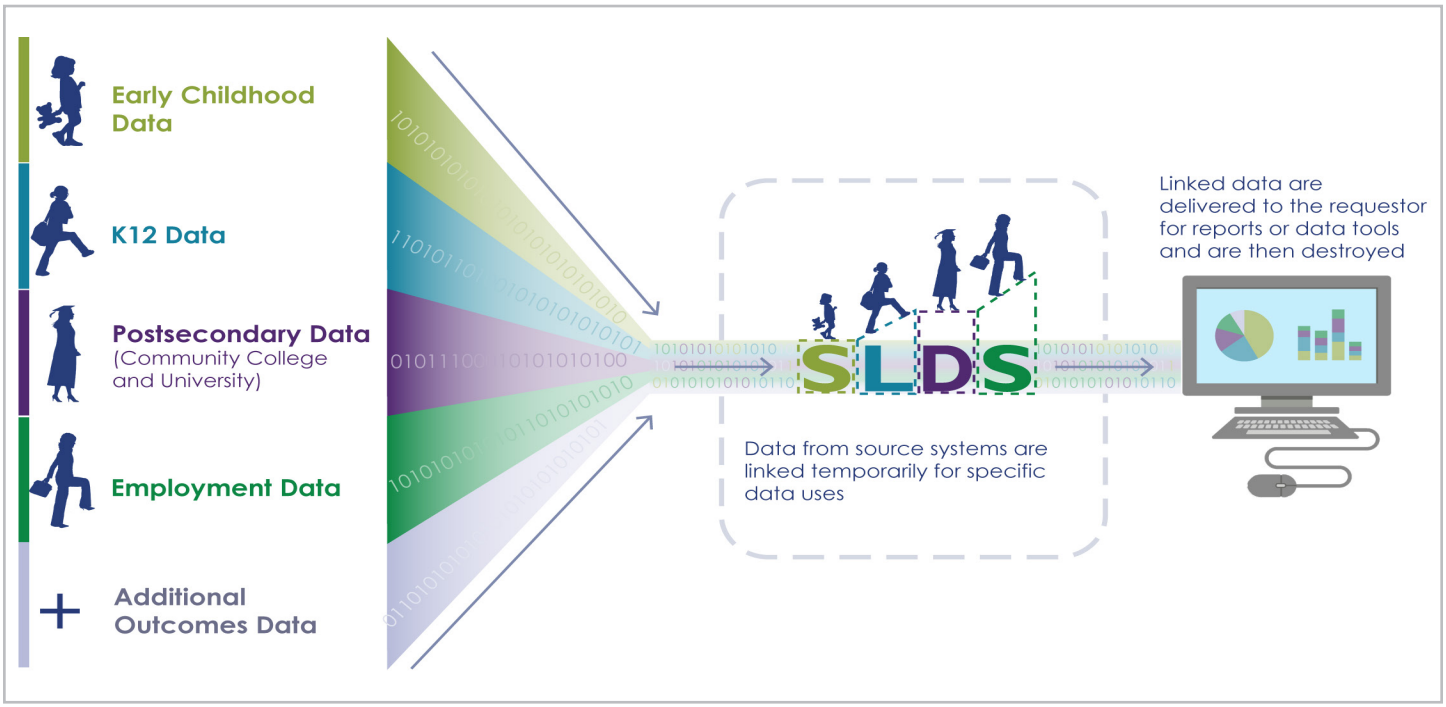

Figure 2. Basic structure of a federated data system

In addition to these two primary system models, many states have implemented a **hybrid model** that combines features of the centralized and federated models to meet states' unique circumstances. Some state agencies may have agreements in place to share data in a common repository, but legal barriers preclude the integration of data from all participating agencies and organizations. For example, postsecondary and workforce agencies might be able to link workforce outcomes data due to common use of Social Security numbers, but the K12 agency might be restricted by statute from using the same common identifier. The structures of a hybrid models vary from state to state, but they commonly maintain some persisting linkages among frequently used data. For example, a hybrid data system might establish links among identifying data (e.g., Social Security numbers, names, dates of birth, and

student identifiers), while keeping separate other types of information, such as enrollment, attainment, and assessment data. These data would be pulled into a dataset only when needed for a specific purpose, such as a report or research request.

Comparison of centralized and federated system characteristics
The following table shows how centralized and federated SLDS models compare on key components of data system design and management. Because hybrid SLDS models can differ greatly from state to state, the hybrid model is not included in this comparison.

Some state agencies may have agreements in place to share data in a common repository, but legal barriers preclude the integration of data from all participating agencies and organizations. For example, postsecondary and workforce agencies might be able to link workforce outcomes data due to common use of Social Security numbers, but the K12 agency might be restricted by statute from using the same common identifier.

Component	Centralized Model	Federated Model
Data ownership	Source agencies own the data and share data stewardship[2] with the centralized data warehouse entity. Data stewardship responsibilities are spelled out in MOUs.	Source agencies own and have stewardship over their data. There is no shared data stewardship.
Staff resources	Staff resources may be required from each source system to oversee and maintain required data access. In addition, support will be needed for the extract, transform, and load (ETL) processes to reflect updates and changes in source data systems and data element modifications. Dedicated or shared staff also will be needed to support the centralized database system.	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will be needed for the extract, transform, and load (ETL) processes to retrieve and link data from participating agency source systems for a data request or report. Staff resources are required from each participating agency to review and approve data requests.
Technical requirements	Each source system will need to have access controls in place for its data or have the technical capability to provide data to the centralized data system. The system infrastructure will require ETL tools as well as capabilities to match and integrate the data, store the results, and deliver integrated datasets to stakeholders via tools such as portals or business intelligence solutions.	Each source system will need to have access controls in place for its data. Source systems also will need hardware and bandwidth to support queries with ETL tools, match the data, and return the resulting dataset, which can then be matched and linked with other source system datasets and delivered to stakeholders via tools such as portals or business intelligence solutions.
System performance	Data extraction is generally fast since needed data matches have already occurred in the transformation and load steps. Data are matched once and used many times. Extracts can be scheduled to occur on source systems during off-peak hours to minimize impact. Centralized data system architecture can be designed specifically to decrease response times.	Data delivery is subject to longer delays due to load and scheduling on each of the individual source systems. Agency-specific performance issues, capacity and other priorities can affect delivery times for requested datasets.
Privacy/security	The centralized data system entity has primary data steward responsibilities, but policies are dictated by source system agencies via MOUs and governance processes. Security is handled through user access controls. The centralized repository may make it easier to preserve data integrity. Although records are typically de-identified during loading, the stakes may be higher in the event of a breach because all data are stored in a single location. However, the risk of breaches may be mitigated by centrally managed security policies and procedures.	Source system agencies have primary responsibility for data privacy and security. Processes are needed for securely handling data queries. Data are diffused, allowing for tailored protection based on the sensitivity of each source system's data and reducing the amount of data that could be accessed through a breach.
Data updates/ corrections	ETL processes take place at specific intervals to capture changes, corrections, and/or updates.	Data reside within each agency. Each agency is responsible for communicating, scheduling and possibly revising the data extraction processes to reflect changes, corrections, and/or updates.

Table 1. Comparison of centralized and federated data systems by key characteristics

2 Data stewardship is a comprehensive approach to data management to ensure quality, integrity, accessibility, and security of the data.

Component	Centralized Model	Federated Model
Data availability	Data must be finalized in the source system and then integrated into the centralized repository before they are available. Access to data is determined by interagency governance processes.	Availability is based on when data are finalized in the source system and ready for extract, as well as how long data are stored. Access to data is determined by the source agency.
Data quality	Consistent data cleansing processes and data quality checks apply to all data as agreed by the contributing source systems. Data may be more reliable and data duplications are more easily identified because they are validated as part of the load process from each source system. Data validation and quality processes may identify and correct data issues in the centralized system, but the source systems are still responsible for correcting their own data.	Quality depends on the validation and other quality processes implemented and supported at each source agency.
Implementation	The initial implementation period is longer due to the need to design and build the centralized database or warehouse. Significant time also is needed to determine infrastructure requirements and establish processes for ETL and data provision.	Implementation is dependent on the capacity and processes in place and supported at each agency.
Scalability	Adding data sources might require supplementing or expanding the centralized data system architecture as well as writing ETL processes and implementing matching and integration rules.	Adding data sources does not require the addition of any hardware or other resources. Adding data sources requires writing ETL processes and implementing matching and integration rules. Each system may be scaled as needed.
Production of standard reports	Standard reports can be automated and scheduled to save time and cost once all the required data are available in the system.	Report production requires one or more agencies to accept this as a responsibility.
Sustainability	Possible approaches to sustainable funding include a state budget appropriation to the centralized data system entity for the development and ongoing support and maintenance of the centralized system. This approach would have no fiscal impact on the participating agencies. Another approach would be for each participating agency to pay a proportional part of the funds needed to support the centralized system in a cost-recovery model managed by the data governance organization. This approach could discourage some agencies from participating in the P-20W+ system.	Possible approaches include asking each agency to contribute resources toward supporting data system processes. This approach could discourage some agencies from participating. Alternatively, state appropriations could be made to each participating agency based on a funding formula to support the data system.
Usability	Longitudinal data are all in one place, facilitating and streamlining data mining.	Data spanning multiple years must be queried from partner agencies, which requires assurance of comparability. If additional years of data are needed for a given cohort, it must be determined whether the additional years of data reside in the source agencies. The entire dataset may need to be rebuilt.
Costs	Initial costs for design, development and implementation of a centralized model is high but ongoing maintenance, hosting, and support can be efficiently managed through using a central organization or trusted third party.	Initial costs are low since the federated system does not require additional infrastructure. Lack of data standards may result in higher costs over time.

Table 1. Comparison of centralized and federated data systems by key characteristics *(continued)*

Key pros and cons to consider

The following table summarizes pros and cons of the centralized and federated data system models.

	Centralized Model	Federated Model
Pros	• Better performance for pulling data • More streamlined for data mining • Easier to account for data integrity/security • Single, central data policy • Easier to ensure data quality • Quicker data results • Avoids issues of disparate and noncompatible technologies	• Shorter time and lower cost for initial implementation • Mitigates turf battles and trust issues • Diffuses data and allows for tailored protection of data based on security • Quicker scalability
Cons	• Higher costs for infrastructure development and training • Data are only as current as the most recent load • Higher risk in the event of a breach due to the amount of data contained in a single repository • More difficult to distribute costs across participating agencies, if needed	• Requires data to be pulled and linked every time a dataset is generated, resulting in delayed results and potentially higher ongoing costs • More difficult to ensure consistent, quality results • Investment and support of intermediary interface by each of the participating agencies • Limited P-20W+ data integration

Table 2. Major pros and cons of centralized and federated data system models

Additional Resources

Building a Centralized P-20W Data Warehouse: SLDS Issue Brief
https://slds.grads360.org/#communities/pdc/documents/3830

Determining Where to House P-20W+ Statewide Longitudinal Data Systems: SLDS Issue Brief
https://slds.grads360.org/#communities/pdc/documents/8635

Interagency Data Governance: Roles and Responsibilities: SLDS Guide
https://slds.grads360.org/#communities/pdc/documents/17093

Linking Early Childhood and K12 Data: A State Example from Kentucky: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/6948

Linking K12 Education Data to Workforce: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5871

Linking K12 Student Data with Postsecondary Data: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5793

P-20W+ Best Practices: SLDS Issue Brief
https://slds.grads360.org/#communities/pdc/documents/5231

Structuring Data for Cross-Sector Longitudinal Reporting: SLDS Issue Brief
https://slds.grads360.org/#communities/pdc/documents/16795

Using DMV Records to Access Social Security Number: SLDS Webinar
https://slds.grads360.org/#communities/pdc/documents/5909

Which ECIDS Model Is Best for Our State ECIDS? SLDS Brief
https://slds.grads360.org/#communities/pdc/documents/6019